Big Earth Science Data Frameworks and Services within NASA's Physical Oceanography DAAC

Presented by David Moroni[1]
(David.F.Moroni@jpl.nasa.gov)

Co-Authors: Suresh Vannan[1], Mike Gangl[1], Ed Armstrong[1], Michelle Gierach[1], Catalina Oaida[2], Jessica Hausman[1]

[1]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA
[2]Raytheon, Pasadena, CA

Presented at ESA LPS 2019

Outline

1. The Role of PO.DAAC
   - Overview of Data Holdings
2. Mission-driven "Big Data" Challenges
3. Legacy Solutions
4. Moving to Cloud:
   a. User-driven Paradigms
   b. Analysis Ready, Interoperable Data
   c. On-demand Analytics
5. Open-source Software/Tools/Services
6. Common Cross-DAAC Interfaces
7. Summary

Role of the Physical Oceanography Distributed Active Archive Center (PO.DAAC) https://podaac.jpl.nasa.gov/
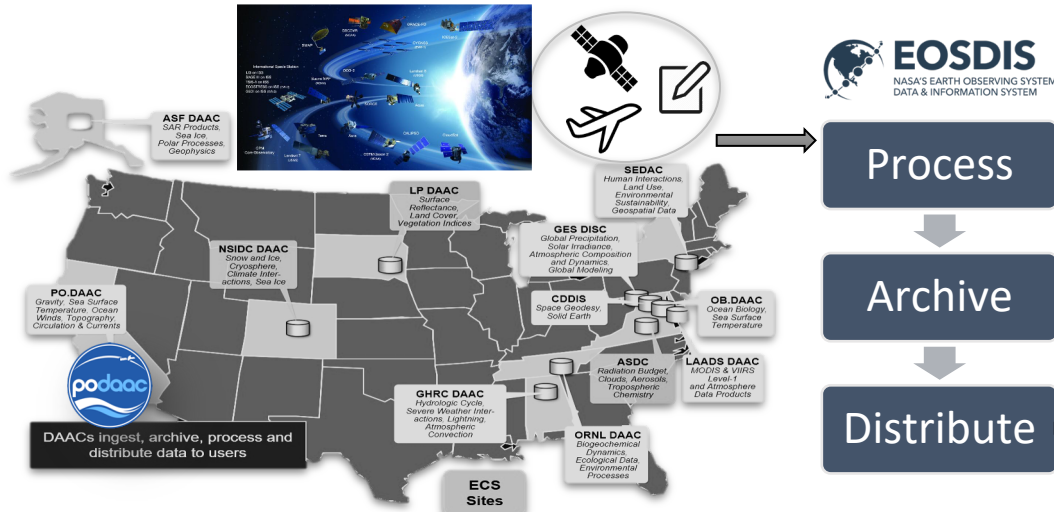
**PO.DAAC Global Data/Services Usage Heatmap**
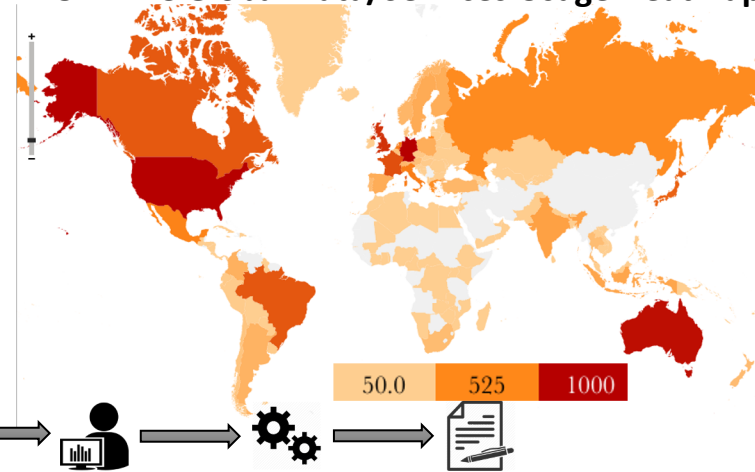
Missions & Projects

Seasat, TOPEX/Poseidon, Jason-1, NSCAT, SeaWinds on ADEOS-II, CYGNSS, GRACE-FO (2018), QuikSCAT, GRACE, GHRSST, MEaSUREs, Aquarius, SPURS, ISS-RapidScat, AirSWOT, OMG, Jason-CS/Sentinel-6 (2020), SWOT (2021)

Parameters

Geodetics/Gravity (85)
Hydrology (11)
Ocean Circulation & Currents (8)
Ocean Surface Salinity (156)
Ocean Surface Topography (6)
Ocean Vector Winds (104)
Sea Surface Temperature (224)
Sea Ice (37)

Tools

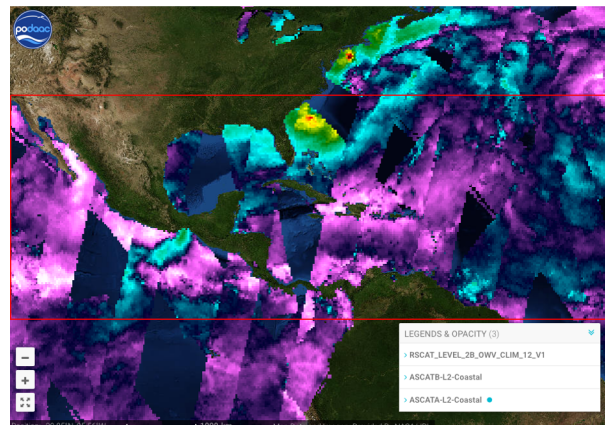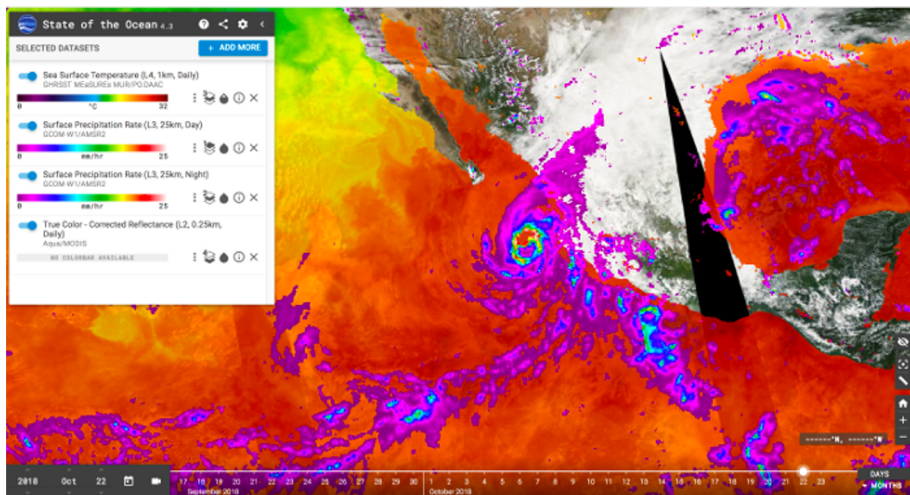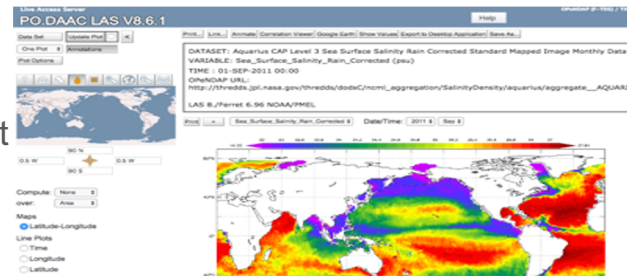Discover
Visualize
Access

3
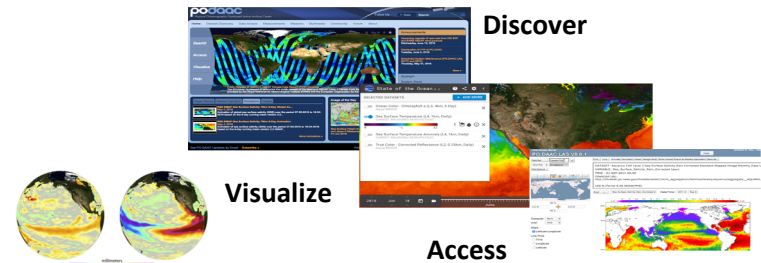
Mission-Driven "Big Data" Challenges

- Lower-cost, more scalable instrument/platform technologies allows for increasingly complex Earth observing missions and data types.
- Tackling the 4 Big V's: **V**olume, **V**elocity, **V**eracity, and **V**ariety.
- **_Volume_** and **_Velocity_**: The Surface Water Ocean Topography (SWOT) Mission, launching Sept. 2021, by itself will deliver a nominal **_50 TB_**/day, with data spikes as high as **_100 TB_**/day.
- **_Veracity_**: checksums to validate data file integrity; automated workflows to validate anticipated latency; metadata validation for completeness and interoperability standards compliance.
- **_Variety_**: many key oceanographic parameters have over 100 unique datasets; Becomes increasingly complex as we branch in to airborne and in situ field campaigns.

4
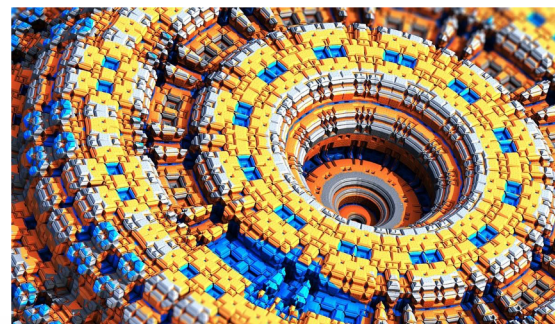
Legacy Solutions (https://podaac.jpl.nasa.gov/dataaccess)

- HTTPS/Web-Dav "Drive" (FTP Replacement)
- OPeNDAP - Subsetting/Transformation
- THREDDS/LAS - Aggregation/Visualization/Analysis/Reduction/Transformation
- State of the Oceans (SOTO) - Multi-layered Visualization/Animation
- Web Services RESTful API - Metadata/Data Query/Reduce/Visualize/Extract
- HiTIDE - Level 2 (i.e., discrete swath geometry) data reduction.

Moving to the Cloud - User-driven Paradigms

- Driven primarily by the Surface Water & Ocean Topography (SWOT) mission.
  - Cloud design architecture incorporates technical interchange with Alaska Satellite Facility (ASF) and ESDIS on the NASA-ISRO Synthetic Aperture Radar (NISAR) Mission.
- SWOT (launch Sep. 2021) will be producing multi-sensor data vital supporting cross-domain science: oceans, surface hydrology, coastal, cryosphere.
- Provides joint collaboration between NASA & CNES, with contributions from UKSA, CSA.
- Analysis-ready formats include: netCDF-4 and GIS/Shapefile
- Extensive user needs assessment and early-adopter program as basis for planning & development.
  - User Needs Assessment, Application Journeys, Use Cases (traceability matrix)
  - Provides input for planning Roadmap and Cloud Services
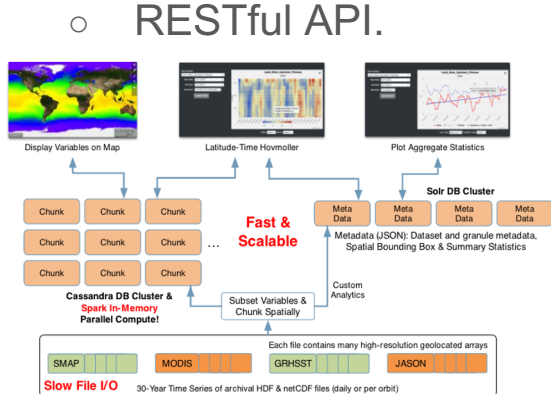 - 3 user communities (Hydrology, Oceans, Coastal)
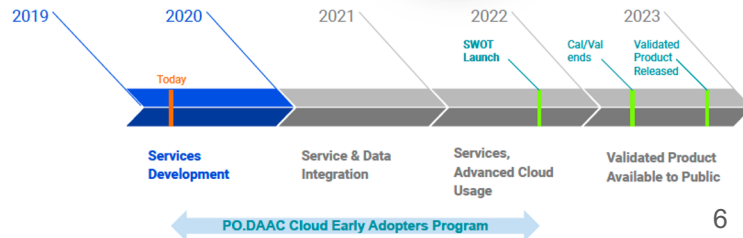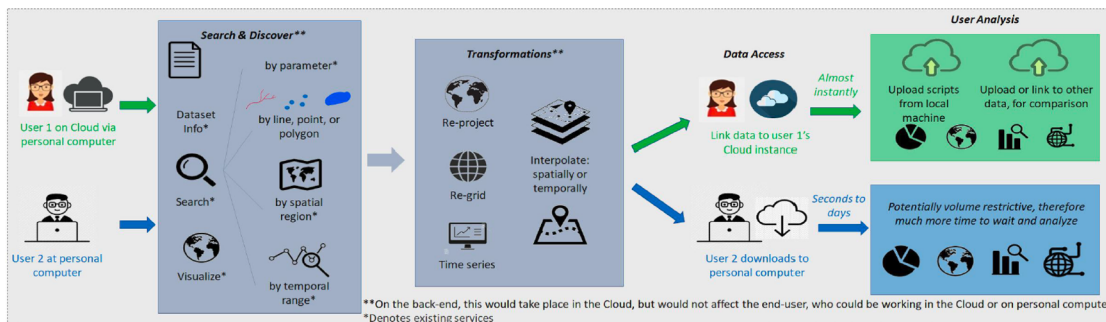
Moving to the Cloud - Analysis Ready and Interoperable

- NASA's Earth Science Data Systems Working Group for Analysis Ready Data (ARD) emphasizes data in a format that is interoperable with "open" assessments and accessibility of data quality (e.g., uncertainty, spatial/temporal characteristics).
- Also makes use of "Analysis-Ready" information as an adjuvant to the physical data.
- CEOS ARD statement: "[ARD is] organized into a form that allows immediate analysis with a minimum of additional user effort and interoperability both through time and with other datasets". http://ceos.org/ard/
- PO.DAAC's best practices supporting ARD via Interoperability: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices
- Interoperability Compliance Validation/Transformation (Leveraging CF, ISO and ACDD): https://podaac-tools.jpl.nasa.gov/mcc/ (Metadata Compliance Checker) https://www.unidata.ucar.edu/software/rosetta/ (Format and Metadata Transformation)
- ARD can also empower remote-access capable Tools/Services, e.g., Panoply, Integrated Data Viewer, THREDDS: https://www.giss.nasa.gov/tools/panoply, https://www.unidata.ucar.edu/software/idv/ https://www.unidata.ucar.edu/software/thredds/current/tds/

Moving to the Cloud - On-demand Analytics

Leveraging NEXUS, open-source technology developed by JPL:

- On-demand Data Discovery, Fusion, Reduction, Visualization, Analytics, Re-packaging.
- Cloud-optimized, horizontal scalability.
- Tiled data storage minimizes I/O bottlenecks.
- Time series, Lat/Lon X Time Hovemoller, Lat/Lon Time Averaging, Correlation Maps.
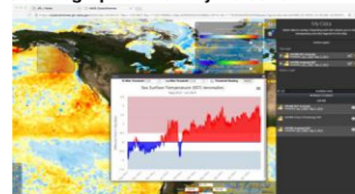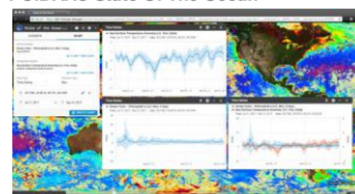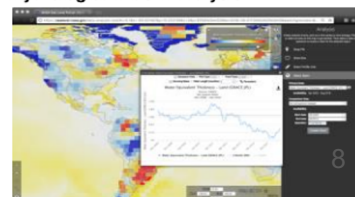- Many more analytics features…
- RESTful API.

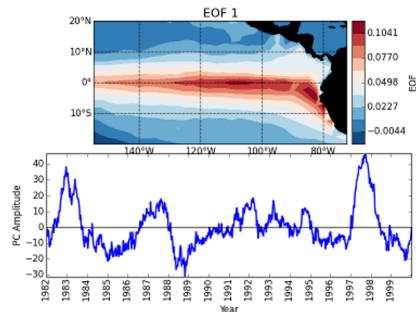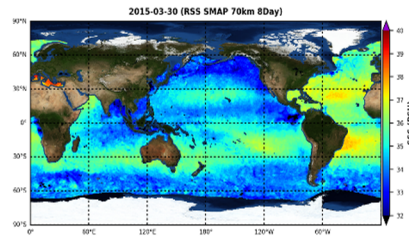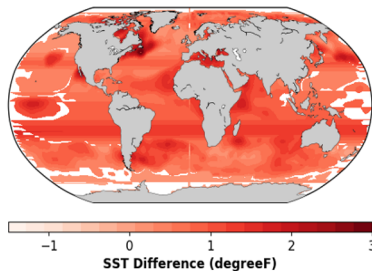Jupyter Notebook - Interactive Workbench
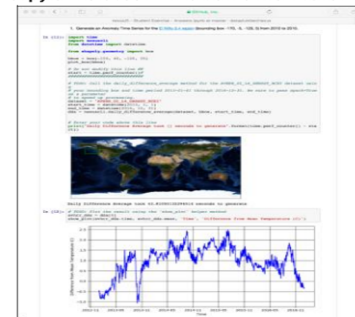
Oceanographic Anomaly Detection

PO.DAAC State Of The Ocean

Hydrological Basin Analysis

NEXUS' Two-Database Architecture

Thomas Huang et al., 2018

Open-source Software/Tools/Services
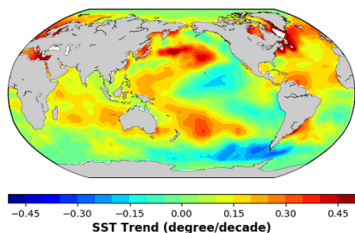
- Data Recipes - https://podaac.jpl.nasa.gov/recipes
  - Reading/Translation (5), Access/Services (29), Visualization (13), Numerical Analysis (6), Tutorials (3)
- PO.DAAC-PY - Python Library Toolkit for PO.DAAC Web Services
  - Updated as of November 2018 - https://github.com/nasa/podaacpy
- NEXUS - On-Demand Visualization, Data Discovery, Re-packaging and Analytics
  - https://github.com/dataplumber/nexus/
  - PO.DAAC GitHub – Coming Soon!

Common Cross-DAAC Interfaces

- NASA Earthdata Code Collaborative (code-sharing for common services).
- NASA's EOSDIS Earthdata Search Client portal (https://search.earthdata.nasa.gov) is already available for all public facing datasets, and will also soon be made available to find and aggregate data from future Big-Data missions across DAACs such as *SWOT* and *NISAR*.
- GIBS/Worldview (https://worldview.earthdata.nasa.gov/): provides multi-layered visualization (over 800 data layers) for all types of satellite-based remote sensing data (historical and current) across all DAACs.
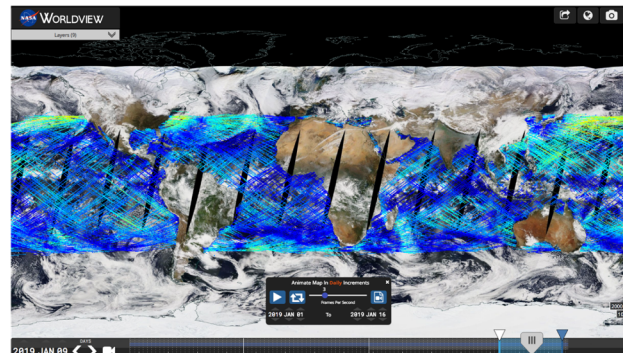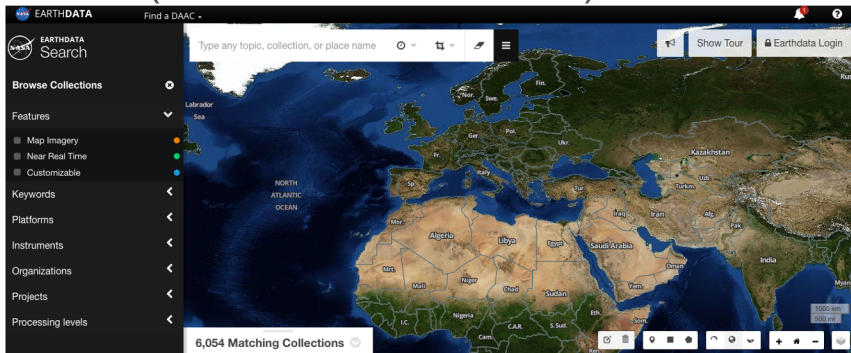
Summary

- 4 V's of Big Data are imposing their will with increasing urgency.
- PO.DAAC has made a critical shift from being reactionary to proactive in anticipating the demands Big Data prior to inundation.
- SWOT is our biggest motivator for moving to cloud data infrastructure.
- NEXUS will continue to increase in maturity and cloud-optimized capabilities in support of a paradigm shift in data usage demand and user needs for more on-demand data-intensive services.
- Open-source solutions will continue growing and evolving, specifically targeting ARD, multi-variate analyses, transformation/re-packaging/reduction.
- PO.DAAC GitHub server is forthcoming, and will be our backbone in our push toward open-source licensing of data-centric technologies and services.
- Cross-DAAC interfaces are key to staying ahead in solution space.
- As we prepare for the cloud, surveying our legacy data holdings for ARD.

Thank you!

PO.DAAC Team and User Working Group - May 2018